AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 108:507–529 (1999)

Testing for Multiple Species in Fossil Samples: An Evaluation and Comparison of Tests for Equal Relative Variation

STEVEN M. DONNELLY[1]* and ANDREW KRAMER[2]
[1]*Department of Anthropology, University of Utah, Salt Lake City, Utah 84112*
[2]*Department of Anthropology, University of Tennessee, Knoxville, Tennessee 37996*

ABSTRACT Tests for equal relative variation are valuable and frequently used tools for evaluating hypotheses about taxonomic heterogeneity in fossil hominids. In this study, Monte Carlo methods and simulated data are used to evaluate and compare 11 tests for equal relative variation. The tests evaluated include CV-based parametric bootstrap tests, modifications of Levene's test, and modified weighted scores tests. The results of these simulations show that a modified version of the weighted scores test developed by Fligner and Killeen ([1976] J. Am. Stat. Assoc. *71:*210–213) is the only test that maintains an acceptable balance of type I and type II errors, even under conditions where all other tests have extraordinarily high type I error rates or little power. Am J Phys Anthropol 108:507–529, 1999.

KEY WORDS species recognition; Levene's test; weighted scores tests; Fligner-Killeen test

The likelihood that multiple species are represented in a collection of fossils is most often evaluated by comparing the magnitude of variation in the fossils with that found in extant species. The latter are assumed to be similar enough to the fossils that they can be used to draw inferences about them (Josephson et al., 1996). Most often they are the extant species that are most closely related to the fossils in question and serve as a measure of how much variation is reasonable for a single species. Because the reference species and the fossils may exhibit differences in absolute variation because they differ in scale, such comparisons are typically based on a measure of relative rather than absolute variation. If comparison of the fossils with the modern analogs shows that the degree of relative variation in the former is greater than can reasonably be expected for a single species, it can be inferred that the fossils represent more than one species.

This would appear to be a straightforward way of testing the hypothesis that a particular group of fossils comprises more than one species. However, determining how much variation is reasonable is complicated by a number of factors. One complicating factor is the nature of the data. Fossil hominid samples are almost always small, sometimes very small, and inevitably these small samples are incomplete and fragmentary. An additional problem, which has received little attention from paleoanthropologists, is that of determining when the degree of relative variation in the fossils is significantly greater than that seen in the reference samples. In this paper, we address the latter problem, using Monte Carlo methods and simulated data to compare and evaluate tests for equal relative variation and taking into account the potential problems created by small sample sizes. The majority of the tests that we evaluate are modified versions of tests that have been developed and

*Correspondence to: Steven M. Donnelly, Department of Anthropology, 102 Stewart Building, University of Utah, Salt Lake City, UT 84112. E-mail: steve.donnelly@healthinsight.org
Received 19 February 1998; accepted 11 November 1998.

used by statisticians to assess absolute variation. We also evaluate several tests for relative variation that have been developed and used by paleoanthropologists to assess hypotheses about taxonomic heterogeneity in fossil hominids.

TESTS AND MEASURES OF RELATIVE VARIATION

Measures of relative variation

To compare variation in fossils and their modern analogs, paleoanthropologists most often rely on measures of relative rather than absolute variation. This is necessary because large organisms tend to vary more than small organisms (Sokal and Rohlf, 1981; Van Valen, 1977). When populations differ appreciably in their means, simple comparisons of their variances or standard deviations can be misleading since differences in absolute variation may reflect nothing more than differences in scale. The statistic that is most often used to quantify relative variation and test hypotheses of multiple species is the coefficient of variation (CV), which is simply the standard deviation expressed as a percentage of the mean $\left(\mathrm{CV} = \frac{s}{\bar{X}} \cdot 100\right)$.

The CV is easy to calculate; it requires only summary statistics which can easily be found in the literature, and it has a simple interpretation. However, there are no simple, robust statistical tests for determining when two CVs are significantly different. Most often the hypothesis that a particular group of fossils represents more than one species is tested by simply comparing sample CVs and determining whether or not the CVs for the fossils exceed those for the reference species (e.g., Cole and Smith, 1987; Grine, 1988; Groves, 1989; Kay, 1982; Kelley, 1986; Kimbel and White, 1988; Kimbel et al., 1985; Martin and Andrews, 1984, 1993; Pan et al., 1989; Stringer, 1986; Teaford et al., 1993; Walker et al., 1993; Wolpoff, 1978). In comparisons of this sort, no attempt is made to determine if the fossil CV is significantly greater than the reference CV. Cope and Lacy (1992, 1994) have developed two versions of a parametric bootstrap test, which are described more fully below, for determining when CVs are significantly different. These tests have recently been used to evaluate hypotheses about variation in the fossils attributed to *Australopithecus africanus* and *A. robustus* (Calcagno et al., 1997; Fuller, 1996; Vinyard, 1997). According to Cope and Lacy (1992, submitted), Monte Carlo simulations using dental metrics from *Cercopithecus* samples that are known to represent single or multiple species show that both tests are capable of detecting multiple species in mixed species samples and have acceptable type I and type II error rates.[1] Neither test, however, has been compared to other tests for equal relative variation. In this study, we compare and evaluate the performance of both tests.

The CV adjusts for differences in variation that are due solely to differences in scale. When it is used to compare variation in two samples for some variable, it is implicitly assumed that the variable in one sample is proportional to that in the other (Lewontin, 1966). Lewontin (1966) pointed out that differences in absolute variation that are due to differences in scale can also be eliminated by taking the logarithms of the variables. The advantage of taking logarithms over using the CV is that it allows for the use of standard statistical tests for homogeneity of variances to determine if there are significant differences between groups in the degree of relative variation.

The range and range-based statistics, such as the range as a percentage of the mean (R%) or the maximum/minimum index (MI), are also sometimes used to quantify relative variation and test hypotheses about variation in fossil samples. According to some authors, the range-based statistics have several important advantages over the CV. Among other things, they are considered to be extremely conservative estimates of variation because they cannot overestimate population variation; they can only underestimate it. Therefore, it is believed that the range-based statistics cannot lead to an incorrect rejection of the null hypothesis of a single species when it is in fact true (Cope, 1993; Martin, 1991; Martin and Andrews,

[1]Type I errors occur when a true null hypothesis is rejected. Type II errors occur when a false null hypothesis is not rejected (Walpole and Myers, 1989).

1993; but see Donnelly, in press). For this reason, it is argued that excessively large values of R% or MI for fossil samples provide very strong evidence that more than one species may be present in a fossil sample. However, the reason range-based statistics are so conservative is that the range is an extraordinarily poor estimator of variation. Among other things, it is extremely sensitive to outliers, and it has long been known that it is both biased and strongly dependent on sample size (Pearson, 1926). The many shortcomings of the range and range-based statistics have recently been demonstrated again by several workers (Cope, 1993; Cope and Lacy, 1995; Foote, 1993). Because the range is such a poor measure of variation, we will not evaluate any tests that use the range or range-based statistics, even though many paleoanthropologists have used and continue to use them as measures of absolute or relative variation (e.g., Coffing and Teaford,1993; Gelvin et al., 1997; Martin and Andrews, 1984, 1993; Miller, 1994, 1995; Miller and Albrecht, 1997; Miller et al., 1998; Pan et al., 1989; Teaford et al., 1993; Walker et al., 1993; Wolpoff, 1978, 1992).

Tests for equal relative variation

Zoologists, paleontologists, and paleoanthropologists have developed a number of statistics and tests for comparing relative variation in two or more samples. Because homogeneity of variances is an important assumption underlying the use of both *t*-tests and analysis of variance, there is also a vast body of statistical literature on the subject of tests for equal variation. For example, Conover et al. (1981) compared and evaluated 56 different tests for homogeneity of variances, and many other tests can easily be found in the statistical literature. All of these could be used as tests of equal relative variation by log-transforming the data.

It is not our intention to evaluate every conceivable statistic and test. Rather, our comparative analyses are limited to 1) tests for equal variation that have been shown to be reasonably robust and powerful[2] and that we believe may be appropriate for the kinds of data and samples that are available to paleoanthropologists and 2) tests that have been recently developed and used by paleoanthropologists for testing hypotheses about taxonomic heterogeneity in fossil samples but whose performance has never been rigorously evaluated.

Most often the hypothesis that a group of fossils represents multiple species is tested by simply comparing sample CVs or other measures of relative variation, such as MI and R%. We do not include these kinds of comparisons in our analyses because our interest is in tests for determining if the differences between groups are statistically significant. In our comparative analyses, we also do not include tests that require large samples, since samples of fossil hominids are often small. Therefore, we do not evaluate tests that require dividing the samples into smaller samples, such as the log-ANOVA test (Martin and Games, 1977). Nor do we evaluate tests that have been shown to lack robustness and/or power or to be deficient in some other respect. For this reason, we do not include any tests that use the range or range-based statistics in our comparative analyses. Based on these criteria we have selected the following 11 tests.

CV ratio test. When the CVs for both samples are less than 30, the ratio of squared CVs can be used as an *F*-test for equal relative variation (Lewontin, 1966). That is, $F = \frac{CV_y^2}{CV_x^2}$, where CV_y and CV_x are the CVs for the fossil and reference samples, respectively. Although this test, which we will refer to as the CV^2 test, is simple and easy to carry out, it has serious limitations. In particular, the *F*-test for equal variances is extremely sensitive to nonnormality (Box, 1953). However, when the data are normally distributed, the *F*-test is asymptotically the most powerful possible test for equal variances (Bailer, 1989).

Cope and Lacy (1992). Cope and Lacy (1992) have developed a parametric boot-

[2]A test is robust when its properties remain reasonably constant in spite of deviations from ideal or assumed conditions (Koopman, 1981). We use the term primarily to refer to the ability of a test to maintain an acceptable type I error rate. *Power* refers to the ability of a test to reject the null hypothesis given that it is false (Myers and Well, 1995).

strap test for equal relative variation that is based on the CV. In this test, which we will refer to as the CL1 test, bootstrapping is used to build a sampling distribution of CVs for samples of size n_y for the reference sample, where n_y is the size of the fossil sample for the variable of interest. The CV at the upper ninety-fifth percentile of this simulated reference distribution is then used as a critical value against which the observed CV for the fossil sample is compared. The test proceeds as follows.

1. Find the CV for the fossil sample.
2. Using the mean and standard deviation for the reference sample, simulate a large, normally distributed population. Cope and Lacy (1992) suggest that this simulated population should be very large. They used an N of 10,000 for their reference populations.
3. Without replacement, draw a random sample of size n_y from this simulated distribution, where n_y is the size of the fossil sample.
4. Calculate the CV for this random sample.
5. Repeat steps 3 and 4 a large number of times, building a sampling distribution of CVs for samples of size n_y. Cope and Lacy (1992) suggest that the number of randomly drawn samples should be very large. In their simulations, they drew 10,000 such samples.
6. Find the CV at the upper ninety-fifth percentile in the sampling distribution. This is the critical value.
7. Compare the critical value with the observed CV for the fossil sample. If the fossil CV exceeds the critical value, reject the null hypothesis.

For our simulations, we have simplified the CL1 test to make it more computationally efficient and less time-consuming. Comparisons of our simplified version of the CL1 test with the original version show that the two tests produce virtually identical results. Typically, rejection rates differ only at the third decimal place. Our version of the CL1 test bypasses the unnecessary step of simulating a large reference population and then sampling from it. Instead, we simply use a random number generator for a normal distribution and the mean and standard deviation of the reference sample to simulate samples of size n_y. In addition, for each test we take only 500 random samples rather than the 10,000 recommended by Cope and Lacy (1992).

Cope and Lacy (1994). Cope and Lacy (1994) have recently developed a more elaborate version of their earlier test. This test, which we will refer to as the CL2 test, attempts to take into account the effect of sampling error on both the reference and fossil CVs. The test proceeds as follows.

1. Calculate the CVs for the reference and fossil samples (CV_y and CV_x, respectively) and find the difference between them, subtracting CV_x from CV_y. This value, d ($=CV_y - CV_x$), is the test statistic.
2. Generate a large simulated reference population. Cope and Lacy (submitted) suggest that whenever possible the reference population should be generated by using the male and female means and standard deviations from the reference sample to simulate separate normally distributed male and female populations and then combining them.
3. Simulate a large, normally distributed population for the fossil sample. This is done by choosing an arbitrary mean and standard deviation such that the CV for the simulated fossil population is equal to that for the simulated reference population. The rationale for this lies in the fact that if the null hypothesis is true the fossil and reference samples are drawn from populations with equal CVs. Because the sex of the fossil specimens is unknown, there is no simple way to simulate the fossil population by combining separate normally distributed male and female populations. Therefore, the fossil population as a whole is assumed to have a normal distribution.
4. Randomly draw samples, without replacement, from each of the simulated populations. These random samples are of the same size as the original samples.
5. Find the CVs for these random samples and calculate d.

6. Repeat steps 4 and 5 a large number of times, building a sampling distribution for *d*.
7. Find the value at the upper ninety-fifth percentile in this sampling distribution. This is the critical value.
8. Compare the critical value with the observed value of *d*. If the observed difference between the CVs is greater than the critical value, reject the null hypothesis.

As noted above, Cope and Lacy (1994) suggest that if possible the simulated reference population should be created by combining separate male and female distributions. Because we are using simulated data, there are no males or females in our samples. However, we use the expedient of dividing our samples on the mean to create approximate male and female samples (cf. Godfrey et al., 1993; Plavcan, 1994). Using the means and standard deviations for the "males" and "females," we then simulate two normally distributed samples, which are combined to form the reference sample. In addition, we use the same simplifications described above for the CL1 test to increase the speed and computational efficiency of the test. Our simplified version of the test produces results that are virtually identical to those obtained for the test as formulated by Cope and Lacy (1994, submitted).

Levene's test. Levene (1960) proposed a test for homogeneity of variances that takes advantage of the fact that a *t*-test or ANOVA for equal means is relatively robust to departures from normality in the underlying distributions. Specifically, Levene (1960) suggested that the raw data be transformed to absolute deviates about their sample means—that is, $X' = |\text{X-mean(X)}|$—and then an ANOVA (or *t*-test if there are only two samples) can be used to determine if there are significant differences between group means for these absolute deviates. The greater the variance within a sample, the greater the group mean for X' will be. Brown and Forsythe (1974) and Conover et al. (1981) both found that the robustness and power of the test are considerably improved when samples are centered on their medians rather than their means. Strictly speaking, this version of Levene's test is a test for equal dispersion (O'Brien, 1978) rather than a test for equal variances because the spread about a median is not a variance (Miller, 1968). In their comparative evaluation of tests for homogeneity of variances, Conover et al. (1981) found that this version of Levene's test was one of the two most robust and powerful of the 56 tests they compared.

Schultz (1983, 1985) showed that Levene's test could be used as a test for equal relative variation if a suitable transformation is used to eliminate the influence of scale. He suggested that this could be achieved by transforming the data to the natural log scale—that is, $X' = \ln X - \text{median}(\ln X)$—or by dividing the absolute deviates about the median by the median—that is, $X' = \dfrac{|X - median(X)|}{median(X)}$. The former is referred to as the mlog and the latter as the mratio test. Schultz (1985) found that both tests are quite robust and powerful for sample sizes greater than 7, but for samples sizes ≤ 7 they become very conservative and lack power. Conover et al. (1981) encountered the same problem with several other tests in which observations are centered on the median. They found, however, that the problem can be alleviated by discarding the median values. Good (1993) suggests that whenever samples are centered on the median the median value should be discarded if the sample *n* is odd or that one of the values bracketing the median should be eliminated if *n* is even.

In our comparative analyses, we evaluate the robustness and power of the mlog and mratio tests as formulated by Schultz (1983, 1985). We also evaluate the performance of these tests when the median or one of the values bracketing the median is discarded. We refer to these as the mlog2 and mratio2 tests.

Talwar and Gentle. Talwar and Gentle (1977) suggested that the power of Levene's test could be improved by using the nonparametric Wilcoxon rank test rather than a *t*-test to determine if there are significant differences between groups with respect to the mean of the absolute deviates, $|X\text{-}mean(X)|$.

In limited simulations, they found that their test, which we will refer to as the TG test, is comparable to Levene's (1960) test for some distributions but more powerful when the underlying distributions have heavy tails.

For our simulations, we use a modified version of the TG test. The data are first transformed to the natural log scale to eliminate any differences in variation that are solely due to differences in scale. Samples are then centered on their medians. Using the median rather than the mean as the measure of central location results in a much more robust test (Conover et al., 1981). In addition, following Good's (1993) suggestion, median values, or one of the values bracketing the median, are discarded to increase the small sample power of the test.

Weighted scores tests for equal dispersion. Mood (1954) was the first to suggest a nonparametric test for equal variances in which the ranks, r_i, of the absolute deviates X - $mean(X)$ are replaced with weighted scores. A number of tests using different weighting schemes have since been proposed, and we evaluate three of these here. For all three tests, the data are first transformed to the natural log scale, samples are centered on their medians rather than their means, and median values or one of the values bracketing the median are discarded.

The first test, which is described more fully by Conover and Iman (1978) and Conover (1980), uses the squared ranks of the absolute deviates, replacing r_i with r_i^2. We will refer to this as the R^2 test. Klotz (1962) proposed using normalized scores, replacing r_i with scores $s_i = \left[\Phi^{-1}\left(\frac{r_i}{N+1}\right)\right]^2$ where Φ is the function for the standard normal distribution and N is the size of the combined sample. Fligner and Killeen (1976) showed that, when sample sizes are small, using scores $s_i = \left[\Phi^{-1}\left(\frac{N+1+r_i}{2(N+1)}\right)\right]^2$ results in a test that is more powerful than the Klotz test. In the comparative evaluations of Conover et al. (1981), the Fligner and Killeen test emerged as one of the two most robust and powerful tests for homogeneity of variances. We evaluate both the Klotz and the Fligner and Killeen tests as tests for relative variation and will refer to the former as the Kl test and the latter as the FK test.

The R^2, Kl, and FK tests differ only with respect to how the ranks are weighted. For each, the test statistic (T) is simply the sum of the scores assigned to the smaller sample, $T = \sum_{i=1}^{n_y}$ where n_y is the size of the smaller sample. When sample sizes are small, critical values of T must be obtained from tables. However, when the sum of the sample ns is greater than 20, which is always the case in our simulations (see below), a large sample approximation $z = \frac{T - E(T)}{\sqrt{Var(T)}}$ can be used (Conover, 1980; Marascuilo and McSweeney, 1977). The expected value of T is simply $E(T) = n_y\bar{s}$, where $\bar{s}$; is the average score, and the variance of T is $Var(T) = \frac{n_y n_n}{(N-1)N}\sum_{i=1}^{N}[s_i - \bar{s}]^2$, where n_y and n_x are the sample sizes for the first and second samples and $N = n_y + n_x$.

METHODS

The robustness and power of the 11 tests described above are evaluated using Monte Carlo methods and simulated data. Because our interest in these tests lies in their potential utility for evaluating hypotheses about taxonomic heterogeneity in fossil samples, the Monte Carlo simulations are intended to mirror the kinds of data and samples that paleoanthropologists must work with. Consequently, the simulations are based on several assumptions. One assumption is that the size of the reference sample (n_x) is always relatively large and as large as or larger than the fossil sample. For our simulated reference samples, n_x is always set at 30. On the other hand, fossil samples are often but not always quite small. It is important to know how well these tests perform over a range of sample sizes for the fossils (n_y) and in particular how well they perform when n_y is small. Therefore, the tests are evaluated with n_y = 5, 7, 10, 15, 22, and 30. In addition, they are evaluated as one-sided tests, because when testing hypotheses of taxonomic heterogeneity the only hypothesis of interest is whether or not variation in the fossil sample is significantly greater than that seen in the comparative samples.

For the simulated reference populations, the CV is always set at 10, with a mean of 10 and a variance of 1.0. For the simulated fossil populations, the means vary randomly between 5 and 20, and the population variances are chosen to give particular values for the population CVs and hence specific values for CV_y^2/CV_x^2. The chosen ratios for the squared CVs are 1.00, 1.86, 2.22, and 2.70. These ratios correspond to "fossil" CVs of 10.0, 13.6, 14.9, and 16.4. As described above, the ratio of squared CVs can be treated as an *F*-ratio and the chosen ratios reflect this; when there are 29 numerator and denominator degrees of freedom (d.f.), $P = 0.05$ for an *F*-ratio of 1.86; *P* also equals 0.05 for an *F*-ratio of 2.22 with 9 numerator and 29 denominator d.f. and for an *F*-ratio of 2.70 with 4 numerator and 29 denominator d.f.

Eight forms of distributions are used to investigate the robustness and power of the tests listed above. These are 1) normal, 2) platykurtotic (heavy-tailed), 3) moderately bimodal, 4) strongly bimodal, 5) moderately leptokurtotic (light-tailed), 6) strongly leptokurtotic, 7) moderately skewed, and 8) strongly skewed. The platykurtotic distributions are uniform distributions. The bimodal distributions are simulated by combining two normal distributions with equal variances but unequal means. Leptokurtotic distributions are simulated by combining two normal distributions with equal means but different variances, and skewed distributions are simulated by using a random number generator for a lognormal distribution.

The 11 tests described above are evaluated on the basis of 500 simulated tests for each combination of sample size, degree of relative variation, and form of the underlying distributions. For all simulated tests, the nominal type I error rate is 0.05, and the proportion of simulated tests that result in rejection of the null hypothesis at $P \leq 0.05$ is used to compare the robustness and power of tests. When F $(= CV_y^2/CV_x^2) = 1.00$, the rejection rate (p_r) provides an estimate of the type I error rate and robustness of a test. When $F > 1.00$, the rejection rate provides an estimate of the power of that test.

The standard error for the rejection rate is largest when $p_r = 0.50$. With 500 tests, the standard error is 0.0224 for $p_r = 0.50$, and for $p_r = 0.05$ (the nominal type I error rate) the standard error of the proportion is 0.0097.

RESULTS

The complete results of the simulations are provided in the appendix. Only the results for the simulations where $F = 1.00$ and 2.70 are given and discussed since the results for the simulations with intermediate values of *F* are entirely consistent with those for the case where $F = 2.70$. Because they provide no additional information, they are not reported or discussed here.

Figure 1 shows the type I error rates when samples are drawn from populations with similar distributions. Conover et al. (1981) define a test as robust if the largest type I error rate is less than 0.10 when the nominal rate is 0.05. By this criterion, the CL1, CV^2, Kl, and FK tests cannot be considered robust. However, the type I error rates of the Kl and FK tests are greater than 0.10 only when $n_y = 5$. Therefore, these tests can also be considered robust provided $n_y \geq 7$. In the case of the CV^2 test, the type I error rate is unacceptably large only when the underlying distributions are leptokurtotic. The CL1 test is by far the least robust. In some cases, the error rate for this test is more than four times the nominal rate, and it is consistently less than 0.10 only when the underlying distributions are platykurtotic or bimodal.

On the basis of the results shown in Figure 1, nine of the 11 tests evaluated can be considered robust if we include the FK and Kl tests for $n_y \geq 7$. Figure 2 shows the power of these nine tests when samples are drawn from populations with similar distributions. The results for all tests are given in the appendix. For the sake of clarity, however, the results for the CV^2 and CL1 tests are not shown in Figure 2 since these tests have unacceptably high type I error rates. The results in Figure 2 show that removing the median value, or one of the values bracketing the median, always increases the power of both versions of Levene's test. When n_y is large, the increase in power is slight, but when n_y is small, removing the median values results in a dramatic in-

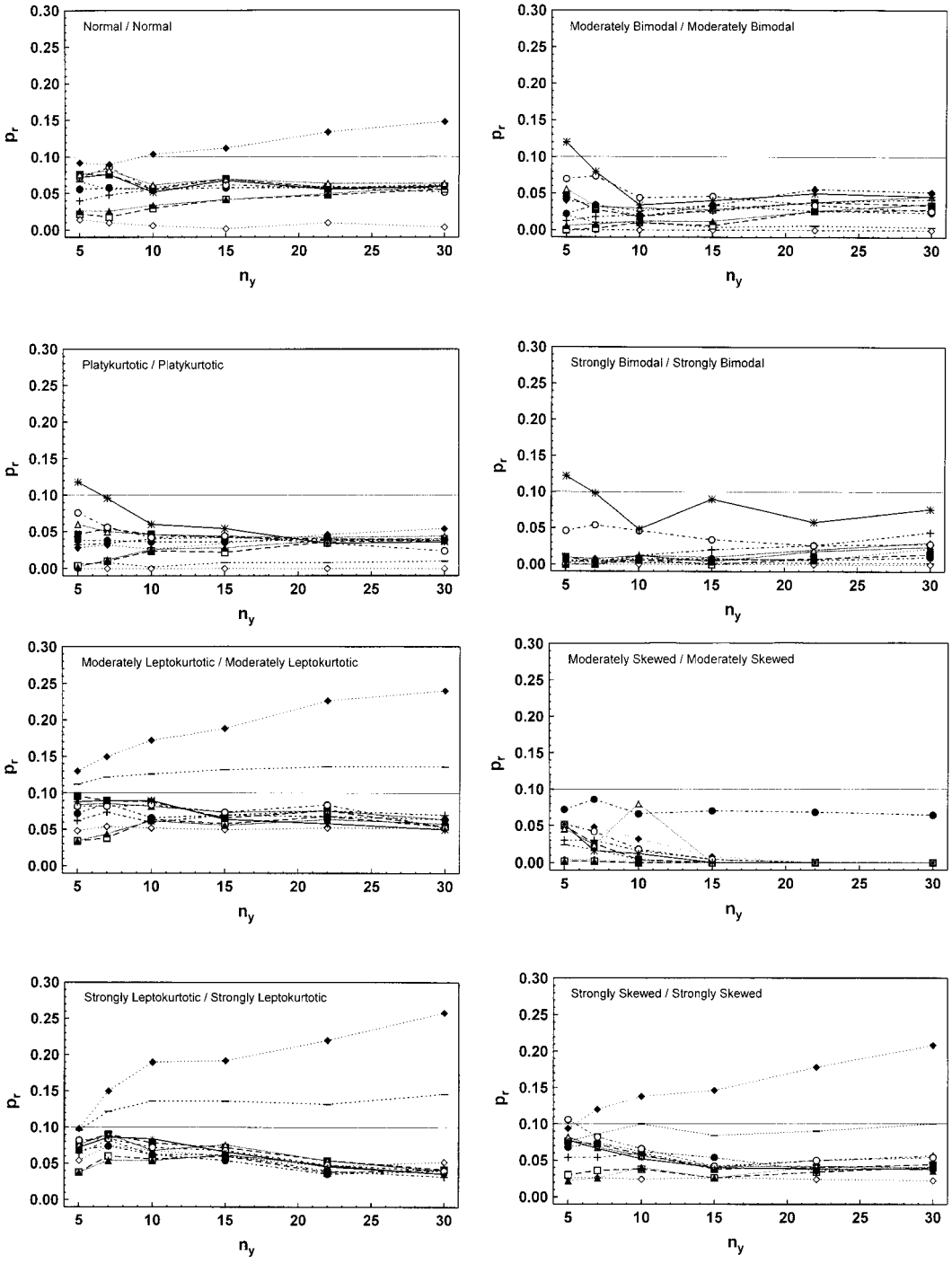

Fig. 1. Results of the simulations showing type I error rates of tests for equal relative variation when samples are drawn from similar distributions. --+--, TG; --●--, R^2; --○--, Kl; --∗--, FK; --□--, mlog; --■--, mlog2; --▲--, mratio; --△--mratio2; ———, CV^2; --♦--, CL1; --◇--, CL2.

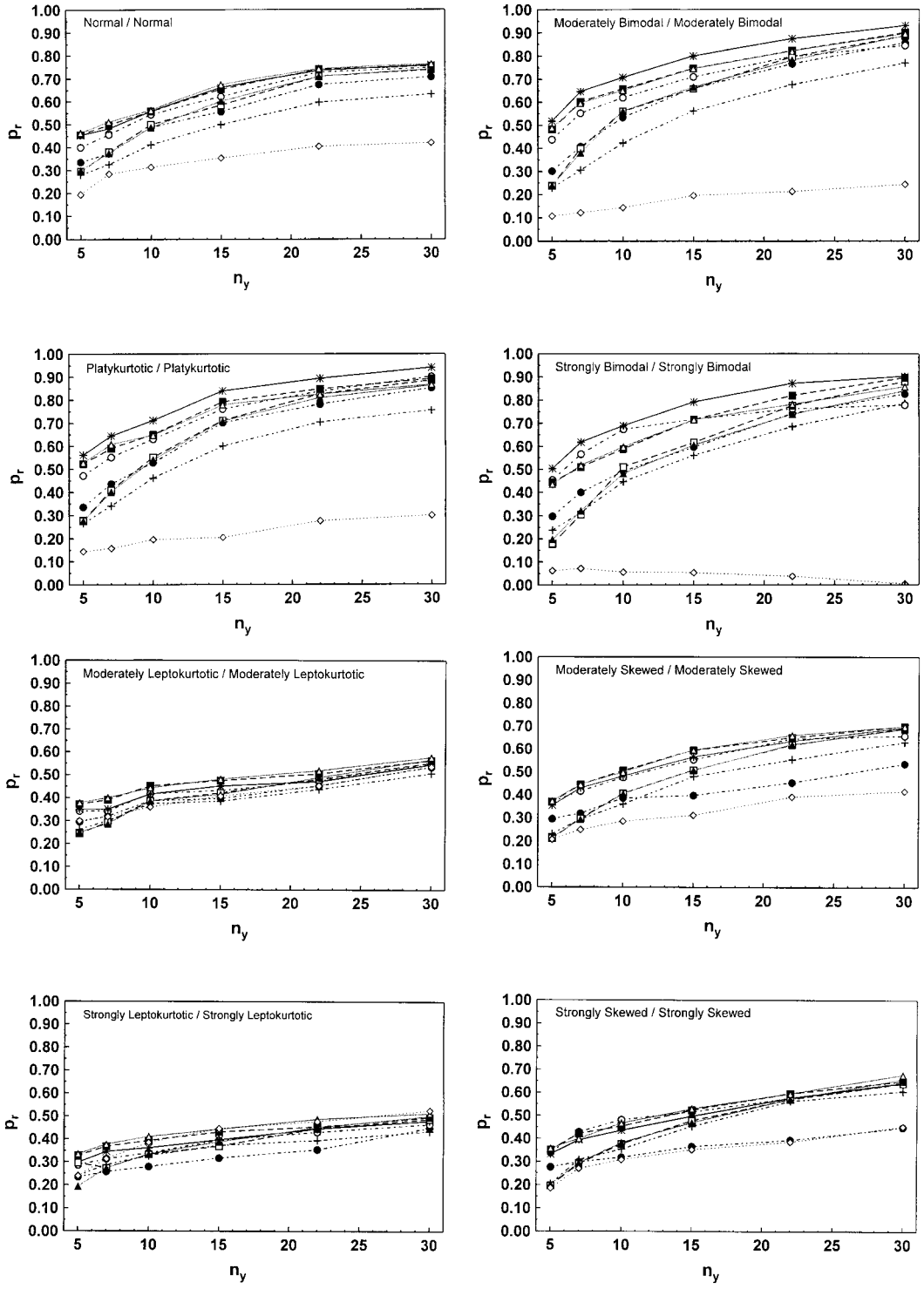

Fig. 2. Results of the simulations showing the power of tests for equal relative variation when samples are drawn from similar distributions. Results for the CV^2 and CL1 tests are not shown. --+--, TG; --●--, R^2; --○--, Kl; --∗--, FK; --□--, mlog; --■--, mlog2; --▲--, mratio; --△--mratio2; --◇--, CL2.

crease in power (cf. Conover et al., 1981). In some cases, for $n_y = 5$ removing the median values more than doubles the power. Because the mlog2 and mratio2 tests are always more powerful than the mlog and mratio tests, we will not consider the latter any further.

The results in Figure 2 also show that overall the tests that consistently have the highest values of p_r are the mratio2, mlog2, Kl, and FK tests, whereas the CL2 test is by far the least powerful. This test is considerably less powerful than the other tests except when the underlying distributions are leptokurtotic. Otherwise, the values of p_r for the CL2 test are often less than half the corresponding values for the most powerful tests, and when the underlying distributions are strongly bimodal the test has virtually no power, with values of p_r that barely exceed and in some cases are actually less than the nominal type I error rate of 0.05. Although it is considerably more powerful than the CL2 test, the TG test also appears to consistently lack power relative to the other tests, particularly for small samples. Like the TG test, the R^2 test lacks power when sample sizes are small, but it is comparable to the FK, Kl, mratio2, and mlog2 tests when n_y is relatively large.

Overall, the results in Figures 1 and 2 suggest that the mratio2, mlog2, and FK tests (for $n_y \geq 7$) all maintain an acceptable type I error rate and are most powerful across a variety of underlying distributions. The Kl test is slightly less powerful, the R^2 and TG tests are even less powerful, especially when n_y is small, and the CL2 test is by far the least powerful. Collectively, the mratio2, mlog2, and FK tests are most powerful when the underlying distributions are platykurtotic or bimodal, and all lose some power when distributions are leptokurtotic or skewed.

The results in Figures 1 and 2 are from simulations in which samples are drawn from populations with similar distributions, but with real data there is no necessary reason this should always be the case. Many of the tests are based on the assumption that the underlying distributions are similar to some degree. To evaluate how well they perform when this assumption is violated, we carried out simulations with six arbitrarily chosen combinations of reference/fossil distributions: strongly skewed/moderately bimodal, normal/strongly bimodal, strongly leptokurtotic/strongly skewed, moderately bimodal/normal, strongly bimodal/moderately skewed, and moderately skewed/moderately bimodal. The results of these simulations for all 11 tests are given in full in the appendix, and the results for selected tests are shown in Figures 3 and 4.

Figure 3 shows the type I error rates when samples are drawn from different distributions for nine of the 11 tests. The results for the CV^2 and CL1 are not shown since these two tests have already been demonstrated to have excessive type I error rates. The results in Figure 3 show that most of the tests that appeared to be quite robust when the underlying populations have similar distributions can have very high error rates when the populations have different distributions. Only the CL2 test maintains an error rate that is always less than 0.10. All other tests have type I error rates that exceed 0.10 for some combinations of distributions, and for all but one of these tests the maximum type I error rate is several times larger than the nominal rate of 0.05. The exception is the FK test. The type I error rate for this test rarely exceeds 0.10, and the test often maintains an acceptable error rate when all other tests except the CL2 test have excessively high rates. Furthermore, the highest type I error rate for the FK test is 0.128. This is considerably less than the maximum values of p_r for the mlog2, mratio2, TG, R^2, and Kl tests, which range from 0.162–0.368.

Of the 11 tests evaluated, only the FK and CL2 tests can be considered robust on the basis of the results shown in Figures 1 and 3. Of these two, only the CL2 test maintains a type I error rate that is always >0.10. However, it is able to do so only because it has very little power, unless one of the populations happens to be leptokurtotic. Figure 4 shows the power of the CL2 and FK tests when samples are drawn from different underlying distributions. These results show that the FK test is always more powerful than the CL2 test. In some cases, the power of the FK test is more than five times greater than that of the CL2 test. Overall,

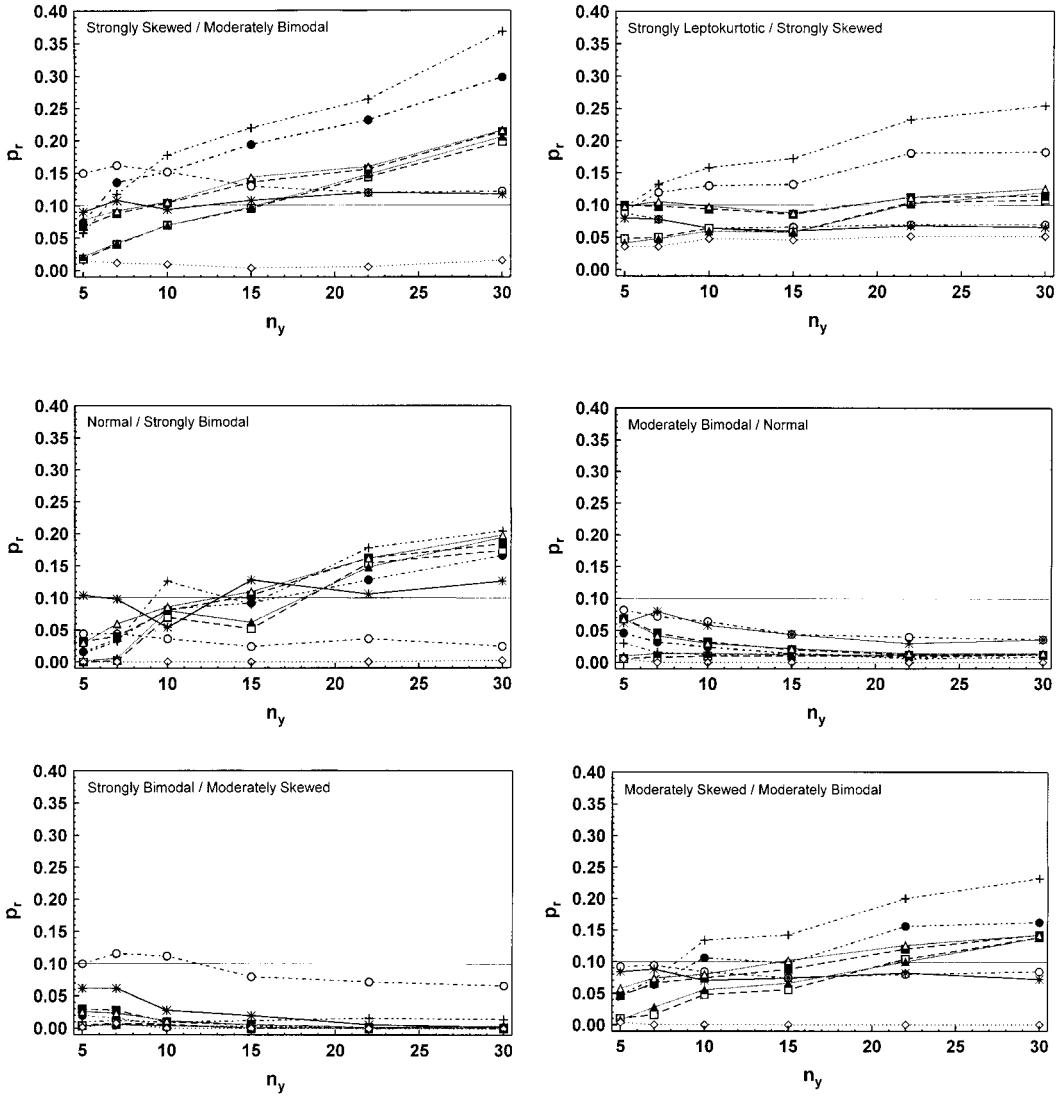

Fig. 3. Results of the simulations showing the type I error rates of tests for equal relative variation when samples are drawn from different distributions. The form of the "fossil" distribution is given first. Results for the CV^2 and CL1 tests are not shown. --+--, TG; --●--, R^2; --○--, Kl; --∗--, FK; --□--, mlog; --■--, mlog2; --▲--, mratio; --△--mratio2; --◇--, CL2.

the FK test provides a much better balance of robustness and power.

DISCUSSION

Comparison of tests

Tests for equal relative variation are often described as though they are capable of directly testing the hypothesis that more than one species is present in a fossil sample. Consequently, comparative analyses of variation in fossils and their extant analogs are often framed as though the null hypothesis that is being directly tested is that there is only a single species represented in the fossil sample and as though the alternative hypothesis that is being tested is that the fossils represent more than one species. Strictly speaking, tests for equal relative

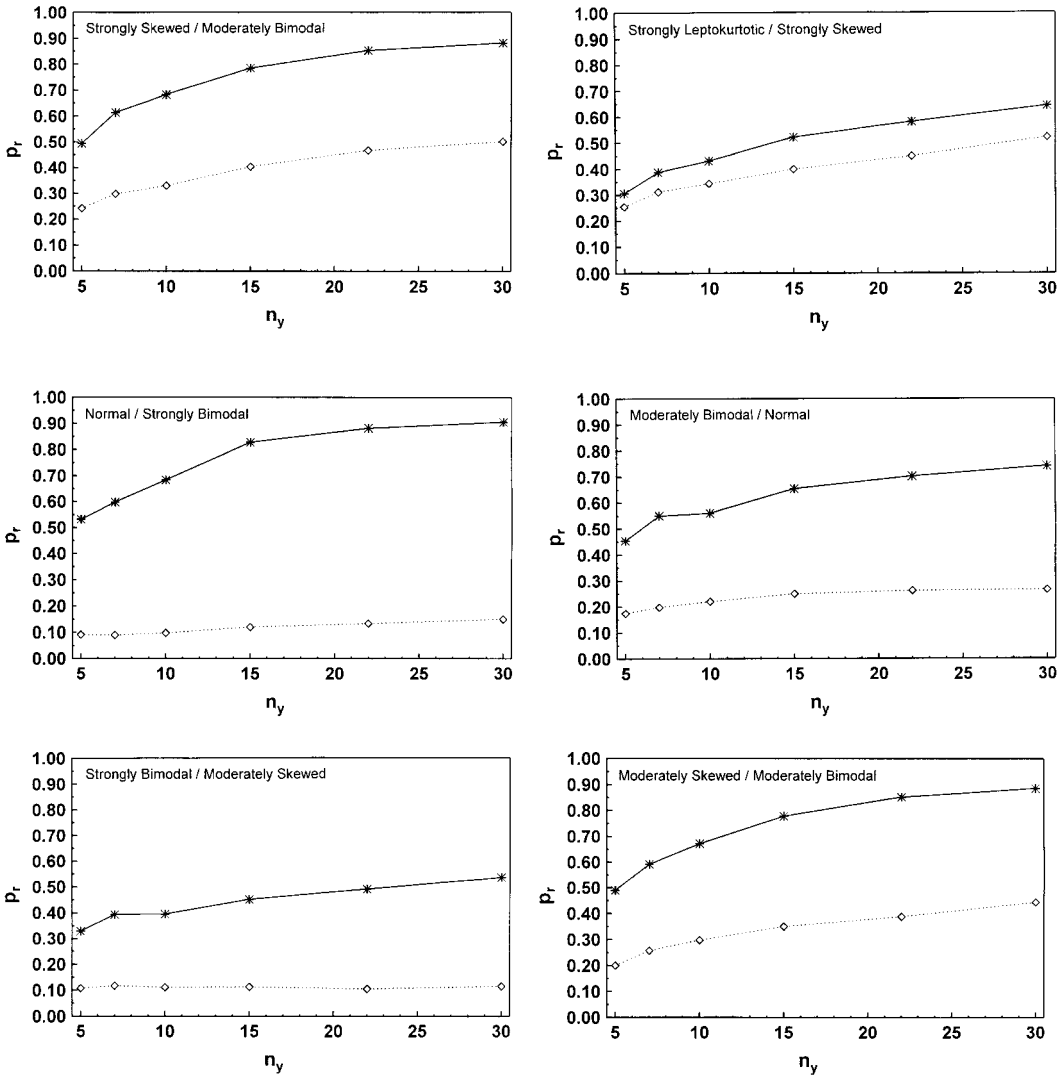

Fig. 4. Results of the simulations showing the power of the FK and CL2 tests when samples are drawn from different distributions. The form of the "fossil" distribution is given first. --∗--, FK; --◇--, CL2.

variation cannot directly test for the presence or absence of multiple species in groups of fossils. The only hypothesis that can be directly tested is whether or not the fossils show more relative variation than the modern comparative species. Martin and Andrews (1993) refer to this as the first-order hypothesis. Just as failure to reject the first-order null hypothesis of equal relative variation does not prove that a particular group of fossils samples only one species (Cope, 1993; Martin and Andrews, 1993; Plavcan, 1993), rejecting the null hypothesis of equal relative variation does not prove that the fossils must represent multiple species. An excessive level of variation in a group of fossils may suggest that multiple species are sampled or provide support for the argument that more than one species is present, but it does not conclusively prove that this must be the case because there are other, alternative explanations that are also possible. As Martin and Andrews (1993) have pointed out, if a test suggests that

there is more variation in a fossil sample than seems reasonable for a single species, there are a number of alternative explanations, which they refer to as second-order hypotheses, that must be considered: 1) the comparative data used are inappropriate; 2) the fossil sample comprises a single species sampled over time so that directional or fluctuating changes in the mean of a trait have inflated the overall level of variation; 3) the sample comprises more than one species; and 4) the sample comprises a single species whose variation at one time exceeds that of the comparative samples employed (cf. Kelley, 1986,1993; Kelley and Xu, 1991). Tests for equal relative variation can only be used to directly test the first-order hypothesis. These tests can tell us if a group of fossils shows an excessive degree of variation; they cannot tell us why the variation appears excessive. By themselves they cannot help us decide which second-order hypotheses may be correct and which may be incorrect. This requires additional evidence and analytical methods.

In spite of this limitation, a test for equal relative variation can be a valuable tool and an important first step in evaluating multiple species hypotheses. However, a test is of little or no value unless we can be confident of its ability to provide us with correct answers. Deciding which of the eleven tests evaluated is best depends upon what criteria are deemed most important. Often the emphasis has been on minimizing the probability of type I errors, and this has been the primary motivation for using range-based measures of relative variation. If the sole criterion for selecting a test is minimizing the type I error rate, then the CL2 test is clearly superior to all other tests. The type I error rate for this test never exceeds 0.10, it is almost always less than the nominal rate of 0.05, and in many cases it is 0.00. Unfortunately, the CL2 test is conservative for the same reason that range-based statistics are conservative: it has very little power. Unless at least one of the distributions happens to be leptokurtotic, the CL2 test is far less powerful and has a much higher type II error rate than any other test. In fact, when the underlying distributions are bimodal, using the CL2 test would make it virtually impossible to reject the null hypothesis. Clearly, a low type I error rate cannot be the only criterion for selecting a test.

In the context of comparing variation in fossils and their extant analogs and evaluating hypotheses of taxonomic heterogeneity, type I and type II errors are both undesirable, and the former are no better or worse than the latter. The optimal test is one that maintains an acceptable type I error rate but is also powerful. By these criteria the FK test, for $n_y \geq 7$, is clearly superior to all other tests. None of the other tests evaluated are comparable to the FK test in terms of their ability to maintain a reasonable type I error rate across a variety of underlying distributions, including situations in which the underlying populations have different distributions. Furthermore, the FK test is always the most powerful or nearly the most powerful test.

The results of the simulations suggest that the CL1 and CL2 tests both have serious limitations. The CL1 test is prone to type I errors, whereas the CL2 test has very little power unless at least one of the populations is leptokurtotic. The high type I error rate for the CL1 test is most likely due to the fact that only the CV for the comparative sample has a sampling distribution, whereas the CV for the fossil sample is treated as though it is a completely accurate measure of variation, known without error, of the population that it represents. The assumption that the comparative population is normally distributed may also contribute to the test's poor performance. The general lack of power of the CL2 test is more difficult to account for. It is most likely due to the assumptions concerning the forms of the fossil and reference populations and the assumption that the fossil population has the same degree of relative variation as the reference sample. This test also appears to be extraordinarily sensitive to the forms of the underlying distributions. When the distributions are strongly bimodal, the test has virtually no power, and when the distributions are normal, skewed, or weakly bimodal it is much less powerful than any other test, but when one or both of the distributions are leptokurtotic it is the most powerful or nearly the most powerful test.

The remaining tests are all roughly comparable in terms of power, with the mlog2, mratio2, and FK tests being most powerful, and all but the CV^2 test maintain an acceptable type I error rate when samples are drawn from similar distributions. All of the tests require some assumptions about the underlying distributions. For some the assumptions are rather specific, whereas for others the assumptions are less restrictive. The CV^2 and Levene's test both assume that the data are normally distributed, the TG test assumes that the populations have similar distributions (Kanji, 1993), and the weighted scores tests assume only that distributions are symmetrical (Conover, 1980). All of these tests except for the CV^2 test appear to be fairly robust to violations of these assumptions, provided that the distributions are similar. However, when the populations have dissimilar distributions, all of these tests, with the exception of the FK test, can have very high type I error rates.

If sample sizes are large enough, diagnostic plots (Chambers et al., 1983; Lee and Tu, 1997) and tests for skewness and kurtosis (Sokal and Rohlf, 1981) can be used to determine how the data are distributed. If the diagnostics indicate that all groups have similar distributions, the mlog2, mratio2, TG, R^2, Kl, or FK tests would be suitable. Any one of these tests could be used without risk of incurring an unacceptably high rate of type I errors. The mratio2, mlog2, or FK test would be the best choice, since these three tests are consistently the most powerful and maintain the best balance between type I and type II errors. If the distributions appear to be similar and there is more than one comparative group, either the mratio2 or the mlog2 test would be the optimal test because they use ANOVA to determine if there are significant differences between groups with respect to the means of X'. Therefore, the multiple comparison procedures that have been developed for ANOVA (e.g., see Hochberg and Tamhane, 1987; Klockars and Sax, 1986; Toothaker, 1993) could be used with the mratio2 or mlog2 test to control and minimize the studywise error rate. In addition, these tests are simpler than the FK test, and, unlike the FK test, they can be carried out with most statistical packages.

When sample sizes are not large, diagnostic plots are difficult to interpret and may even be misleading, and tests for skewness and kurtosis will have little power. Since hominid and primate fossil samples are often small, there will be many instances in which it will be difficult or impossible to determine how the fossil data are distributed. When this occurs, it may seem reasonable to simply assume that the underlying distributions are similar. With real data, however, there is no justification for making such an assumption. In fact, there are many situations in which we would expect the distributions to be different. Simply assuming that the distributions are similar could easily lead to an inappropriate test and an unacceptably high risk of type I errors.

To determine how likely it is that different distributions would be encountered when comparing fossil hominids with their modern analogs, we have looked at distributions for the mesiodistal (m-d) and buccolingual crown diameters for the mandibular molars and premolars of two species of modern hominoids, *Pan troglodytes* and *Gorilla gorilla* (data from Mahler, 1973) and two species of fossil hominids, *A. afarensis* (data from Johanson et al., 1982; White, 1977, 1980) and *A. robustus* (data from Grine and Daegling, 1993; Grine and Strait, 1994; Wood, 1991). For these data, there are very few variables for which all four groups have similar distributions, and tests for relative variation across the four taxa would entail many comparisons of samples drawn from dissimilar distributions. For example, as shown in Figure 5, for M_2 m-d the gorilla distribution ($n = 44$) is leptokurtotic, the chimpanzee distribution ($n = 34$) is moderately skewed to the left, the *A. afarensis* distribution ($n = 18$) is strongly platykurtotic and appears to be nearly uniform, and the *A. robustus* distribution ($n = 19$) is moderately skewed to the right and somewhat platykurtotic. The sample sizes for *A. afarensis* and *A. robustus* are relatively large for fossil hominids but marginal for diagnostic plots. These samples may actually be derived from platykurtotic and skewed distributions, or they may only appear to be

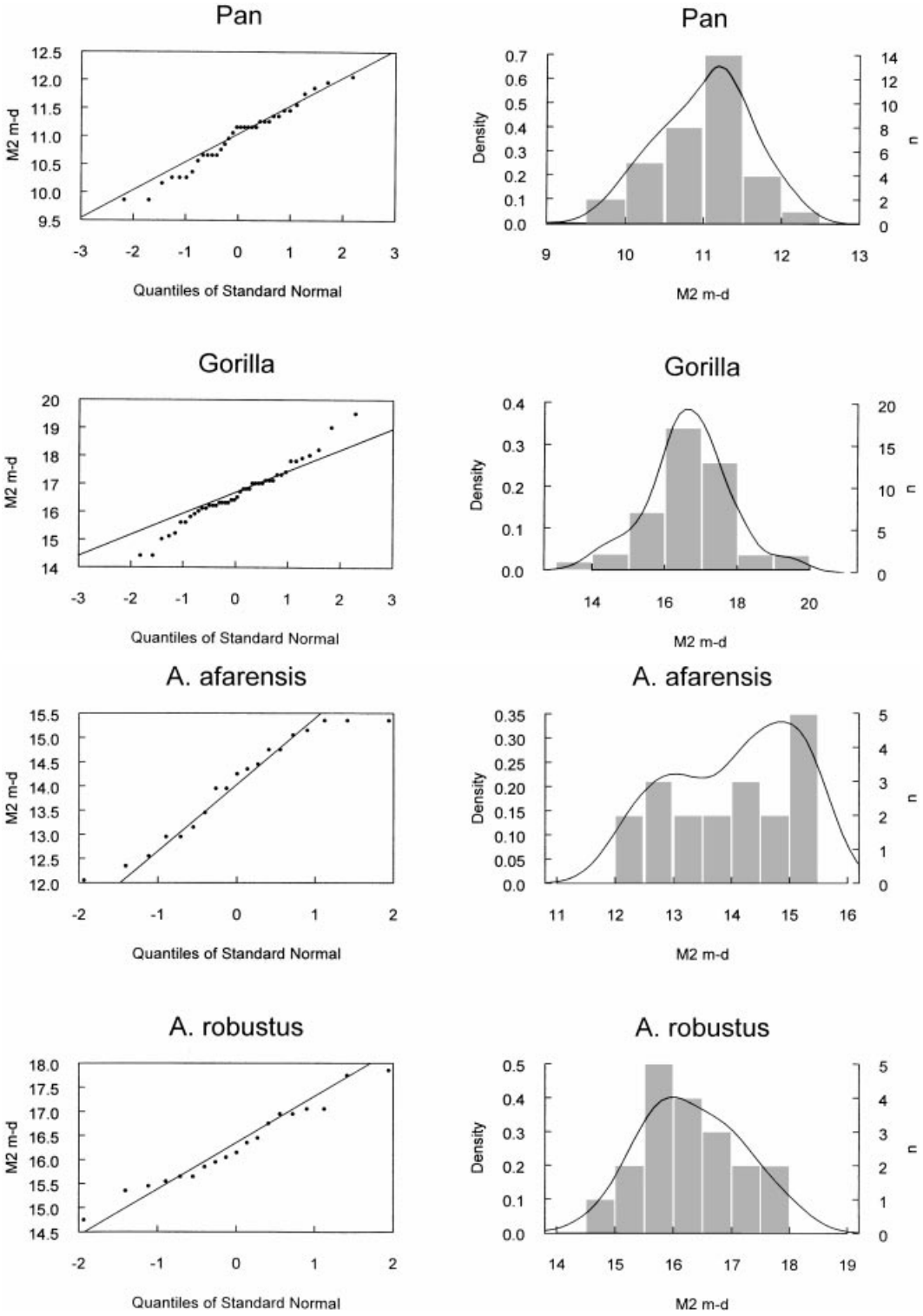

Fig. 5. Quantile-quantile plots and frequency histograms with smoothed densities superimposed showing the distributions for M_2 m-d for *Pan troglodytes* ($n = 34$), *Gorilla gorilla* ($n = 44$), *Australopithecus afarensis* ($n = 18$), and *Australopithecus robustus* ($n = 19$).

derived from such populations because of sampling error. With the marginal sample sizes, there is no way to be certain which alternative is correct. Undoubtedly, though, the *Pan* and *Gorilla* samples are derived from different distributions, and the available evidence suggests that the fossil samples are also derived from distributions that are different from one another and from those of the modern hominoids. For M_2 m-d, all of the tests except for the FK test or the CL2 test would be subject to an unacceptably high type I error rate because of the differences in the distributions. This particular variable was chosen because it is an extreme case, with all four groups having different distributions; however, as noted above, we found very few measurements for which all four taxa appear to have similar distributions. Clearly, there is no justification for simply assuming that all samples are derived from similar distributions.

If diagnostic plots indicate that there are differences in the forms of the distributions, or if the sample sizes are so small that diagnostic plots and tests cannot be used, then the FK test is the only test that should be used. The results of the simulations with different distributions show that tests that are robust when distributions are similar can have very high type I error rates when the distributions are different. The mlog2, mratio2, TG, Kl, and R^2 tests all have type I error rates that exceed 0.10 for some combinations of different distributions, with maximum values of p_r that may be several times greater than the nominal error rate of 0.05. In contrast, when distributions differ, the type I error rate for the FK test rarely exceeds 0.10, and its maximum error rate is only slightly greater than 0.10 and considerably lower than that of any other test. When distributions are different or when sample sizes are so small that the forms of the distributions cannot be reliably determined, the FK test is the only appropriate test. It is the only test that maintains a reasonable balance between type I and type II errors under these conditions. However, it should only be used when sample sizes are ≥ 7, since for sample sizes smaller than 7 it is somewhat more prone to type I errors.

An example

To provide a more explicit description of how the Fligner and Killeen test can be used as a test for equal relative variation, we present an example using real data. We will use the test to compare the degree of relative variation for P_4 m-d in *A. afarensis* ($n_y = 15$) (data from Johanson et al., 1982; White 1977, 1980) with that seen in *G. gorilla* ($n_x = 24$) (data from Mahler,1973). Cole and Smith (1987) found that, relative to modern hominoids and other fossil hominids, *A. afarensis* seems to show an unusually large degree of relative variation for this variable. For our *A. afarensis* sample, the CV is 10.45 (95% CI, 7.63–16.69), whereas for *G. gorilla*, which has the highest CV of any extant pongid for this variable (unpublished data), the CV is only 7.19 (95% CI, 5.59–10.12).

The raw data for both samples are given in Table 1, along with the values of x_i' (= $|\ln(x_i) - \text{median } \ln(x_i)|$), the ranks of x_i' (r_i), and the normalized scores (s_i). After the x_is are log-transformed and converted to absolute deviates about the sample medians, the median (if the sample n is odd) or one of the values bracketing the median (if the sample n is even) is discarded in each sample. In the example given here, AL 333w-32, which is the median for *A. afarensis*, is dropped, and observation 12, which is one of the values bracketing the median of the gorilla sample, is also dropped. The two samples are then combined and the x_i' ranked. Thus, for example, AL 145–3, which has the smallest value of x_i' (0.00), receives a rank of 1, and AL 128–2, which has the largest value (0.2101) is given a rank of 37. Note that ties among the x_i's are not resolved at this point, and tied values are given sequential rather than average ranks. For example, $x_i' = 0.1735$ for both LH 3 and LH 14, but they are given ranks of 31 and 32, respectively, rather than both receiving the average rank of 31.5.

The values of r_i are subsequently converted to weighted scores using the equation $s_i = \left[\Phi^{-1}\left(\frac{N + 1 + r_i}{2(N + 1)}\right)\right]^2$. This is done by finding the quantile of the standard normal distribution that corresponds to the term within the inner brackets and then squaring it. For example, for AL 145–3, which has a

rank of 1, the term within the inner brackets evaluates to 0.513. The quantile of the standard normal distribution that corresponds to this is 0.033, which equals 0.001 when squared. For AL 128–2, which has a rank of 37, the term in the inner brackets equals 0.987. The corresponding quantile of the normal distribution is 2.222, and $s_i = 4.935$. After the ranks are converted to weighted scores, ties are resolved by averaging the s_is for the tied values. For example, $s_i = 1.879$ for LH 3 and LH 14. This is the average of the weighted scores for $r_i = 31$ ($s_i = 1.763$) and 32 ($s_i = 1.994$).

The test statistic, *T*, is the sum of normalized scores for the smaller sample, which in this case is *A. afarensis*, for which $\Sigma_{i=1}^{n_y} s_i = 19.474$. To determine if *T* is statistically significant using the large sample approximation $z = \dfrac{T - E(T)}{\sqrt{Var(T)}}$, we must also find the expected value and the variance of *T*, which are equal to $n_y\bar{s}$ and $\dfrac{n_y n_x}{(N-1)N}\Sigma_{i=1}^{N}[s_i - \bar{s}]^2$, respectively. In our example, $\bar{s} = 0.855$, which makes the expected value of $T = 12.826$ and the variance of $T = 11.635$. Inserting these values into the equation for the large sample approximation, we find that $z = 1.949$. With a one-sided alternative, $P = 0.0256$. Thus, the FK test demonstrates that for P_4 m-d *A. afarensis* shows significantly greater relative variation than *G. gorilla*.

TABLE 1. Raw data, transformed values, ranks, and weighted normalized scores for the Fligner-Killeen test for Australopithecus afarensis *and* Gorilla gorilla *for P_4 m-d*[1]

Specimen	x_i	x_i'	r_i	s_i
Australopithecus afarensis ($n = 15$)				
AL 128-2	7.7	0.2101	37	4.935
AL 288-1i	8.2	0.1472	34	2.624
AL 417-1	8.6	0.0995	28	1.252
AL 198-1	8.9	0.0652	22	0.647
AL 207-13	9.2	0.0321	12	0.165
AL 333w-1	9.4	0.0106	2	0.004
AL 145-3	9.5	0.0000	1	0.001
AL 333w-32	9.5	—	—	—
AL 266-1	9.7	0.0208	6	0.040
AL 400-1a	9.9	0.0412	15	0.267
LH 4	10.3	0.0808	25	0.902
AL 277-1	10.4	0.0905	27	1.122
LH3	10.9	0.1375	31	1.879
LH 14	10.9	0.1375	32	1.879
AL 333-44	11.1	0.1556	36	3.756
Gorilla gorilla ($n = 24$)				
1	9.2	0.1320	29	1.399
2	9.8	0.0688	23	0.724
3	9.9	0.0587	18	0.401
4	10.1	0.0387	13	0.210
5	10.1	0.0387	14	0.210
6	10.2	0.0288	8	0.103
7	10.2	0.0288	9	0.103
8	10.2	0.0288	10	0.103
9	10.2	0.0288	11	0.103
10	10.3	0.0190	4	0.022
11	10.3	0.0190	5	0.022
12	10.3	—	—	—
13	10.7	0.0190	3	0.010
14	10.8	0.0283	7	0.054
15	11.0	0.0467	16	0.307
16	11.1	0.0557	17	0.353
17	11.2	0.0647	19	0.515
18	11.2	0.0647	20	0.515
19	11.2	0.0647	21	0.515
20	11.3	0.0736	24	0.809
21	11.4	0.0824	26	1.006
22	12.0	0.1337	30	1.568
23	12.1	0.1420	33	2.274
24	12.2	0.1502	35	3.086

[1] r_i, rank of x_i'; $s_i = \left[\Phi^{-1}\left(\dfrac{N+1+r_i}{2(N+1)}\right)\right]^2$; x_i, raw data (mm); x_i', $|\ln(x_i) - \text{median}\ln(x_i)|$.

SUMMARY AND CONCLUSIONS

Tests for equal relative variation can be important and valuable tools for evaluating hypotheses about variation and taxonomic heterogeneity in fossil samples. However, they are of little or no value unless we can be confident of their ability to give correct answers. A poor choice of test can lead to erroneous conclusions and is worse than no test at all, since it can give a false sense of security in the accuracy and validity of the results. The procedure that has been used most often to test multiple species hypotheses is to simply compare sample CVs or other measures of relative variation, such as R% or MI. With this approach, no attempt is made to take into account the possible effects of sampling error on the sample statistics. Simple comparisons of this sort should be avoided because they have extraordinarily high type I error rates, even for range-based statistics (Donnelly, submitted). When we evaluate hypotheses about variation in fossil samples, the objective should always be to determine if there are statistically significant differences between groups in the degree of relative variation.

In the present study, Monte Carlo methods and simulated data have been used to evaluate and compare 11 tests for equal

relative variation. The tests evaluated include the CV-based, parametric bootstrap tests developed and used by Cope and Lacy and modified versions of Levene's test, the Talwar and Gentle test, and weighted scores tests. The results of these simulations show that the parametric bootstrap tests have serious limitations. The CL1 test has an extremely high type I error rate, whereas the CL2 test typically has little power and a high type II error rate. The different versions of Levene's test, the TG test, and the weighted scores tests all maintain acceptable type I error rates and are reasonably powerful for a variety of underlying distributions, as long as the distributions are similar in form. However, all but the FK test can have very high type I error rates when the underlying distributions differ. Our modified version of the Fligner and Killeen test is the only test for relative variation that maintains an acceptable balance between type I and type II errors when populations have different distributions. If diagnostic plots and tests indicate that samples are derived from different distributions or if sample sizes are so small that diagnostic plots and tests cannot be used, then the FK test is the only appropriate test. However, it should be used only when sample sizes are ≥ 7. This restriction effectively sets a lower limit of 7 for the size of the fossil samples that can be used in tests for relative variation.

ACKNOWLEDGMENTS

We thank Dana Cope for several helpful discussions on this topic and for allowing us access to his unpublished manuscript. In addition, we are grateful to Alan Rogers, Lyle Konigsberg, Leigh Van Valen, two anonymous reviewers, and especially Sharon Donnelly for their advice, comments, and criticism.

APPENDIX. Results of the simulations[1]

Distribution	F	n_y	CV[2]	CL1	CL2	mlog	mlog2	mratio	mratio2	TG	R^2	Kl	FK
Normal/normal	1.00	5	.066	.092	.014	.022	.076	.026	.076	.040	.056	.072	.072
		7	.056	.090	.010	.018	.076	.026	.082	.048	.058	.086	.076
		10	.058	.104	.006	.030	.056	.034	.062	.056	.056	.056	.052
		15	.070	.112	.002	.042	.070	.042	.070	.058	.058	.062	.068
		22	.058	.134	.010	.048	.058	.050	.064	.058	.060	.058	.056
		30	.058	.148	.004	.056	.060	.060	.064	.062	.056	.052	.060
	2.70	5	.444	.482	.194	.296	.456	.300	.464	.280	.336	.400	.454
		7	.520	.570	.284	.380	.498	.376	.512	.326	.374	.456	.482
		10	.602	.660	.314	.500	.558	.488	.564	.412	.488	.544	.558
		15	.718	.820	.354	.584	.654	.602	.676	.500	.556	.622	.662
		22	.810	.902	.404	.710	.740	.710	.744	.596	.674	.728	.736
		30	.828	.924	.420	.738	.754	.742	.764	.632	.708	.748	.758
Platykurtotic/platykurtotic	1.00	5	.006	.028	.000	.004	.046	.002	.060	.032	.038	.076	.118
		7	.008	.032	.000	.010	.054	.012	.050	.034	.038	.056	.096
		10	.002	.026	.000	.024	.046	.026	.046	.040	.036	.042	.060
		15	.008	.034	.000	.022	.042	.028	.044	.042	.036	.044	.054
		22	.008	.046	.000	.036	.040	.038	.042	.038	.040	.034	.034
		30	.010	.054	.000	.038	.040	.040	.044	.038	.038	.024	.036
	2.70	5	.480	.532	.144	.278	.524	.276	.528	.266	.336	.472	.562
		7	.556	.640	.158	.408	.590	.402	.608	.342	.436	.552	.644
		10	.680	.766	.196	.552	.650	.542	.648	.462	.528	.630	.712
		15	.784	.886	.205	.712	.792	.710	.780	.600	.702	.762	.840
		22	.866	.940	.276	.826	.848	.810	.826	.704	.782	.838	.894
		30	.914	.974	.300	.886	.892	.864	.868	.756	.852	.902	.942
Moderately bimodal/ moderately bimodal	1.00	5	.020	.040	.000	.000	.046	.006	.056	.012	.022	.070	.120
		7	.010	.034	.000	.002	.028	.008	.032	.018	.034	.074	.080
		10	.010	.026	.000	.010	.018	.012	.030	.020	.018	.044	.034
		15	.004	.034	.000	.006	.028	.012	.028	.026	.034	.046	.040
		22	.006	.056	.000	.026	.038	.026	.038	.038	.026	.034	.050
		30	.004	.052	.000	.028	.034	.036	.046	.042	.024	.026	.046
	2.70	5	.480	.550	.108	.238	.482	.240	.490	.230	.302	.438	.518
		7	.600	.688	.122	.400	.602	.382	.596	.306	.408	.552	.646
		10	.684	.798	.144	.560	.658	.562	.652	.424	.534	.620	.708
		15	.780	.896	.196	.658	.746	.666	.746	.562	.660	.710	.800
		22	.848	.968	.214	.796	.822	.782	.822	.676	.766	.798	.874
		30	.924	.990	.244	.886	.900	.888	.896	.770	.856	.844	.930
Strongly bimodal/ strongly bimodal	1.00	5	.002	.004	.000	.000	.010	.000	.010	.004	.006	.046	.122
		7	.004	.008	.000	.000	.004	.000	.008	.004	.004	.054	.098
		10	.002	.002	.000	.006	.006	.012	.012	.012	.008	.046	.048
		15	.002	.008	.000	.000	.006	.004	.010	.020	.008	.034	.090
		22	.002	.006	.000	.008	.008	.018	.020	.026	.006	.026	.058
		30	.002	.022	.000	.014	.014	.024	.030	.044	.010	.028	.076
	2.70	5	.496	.562	.062	.178	.444	.198	.436	.238	.298	.454	.504
		7	.596	.682	.072	.306	.510	.322	.518	.314	.400	.566	.618
		10	.708	.816	.056	.510	.588	.482	.598	.446	.492	.674	.690
		15	.830	.920	.054	.616	.716	.606	.716	.560	.596	.718	.792
		22	.912	.980	.038	.776	.820	.740	.782	.686	.742	.762	.872
		30	.956	.990	.004	.880	.896	.840	.858	.786	.826	.778	.902
Moderately leptokurtotic/ moderately leptokurtotic	1.00	5	.112	.130	.048	.034	.096	.034	.084	.062	.072	.082	.088
		7	.122	.150	.054	.038	.090	.044	.086	.074	.086	.082	.090
		10	.126	.172	.052	.064	.088	.062	.082	.060	.066	.084	.090
		15	.132	.188	.050	.058	.066	.056	.074	.070	.070	.074	.064
		22	.136	.226	.052	.064	.076	.064	.076	.076	.068	.084	.058
		30	.136	.240	.052	.054	.054	.058	.064	.070	.064	.052	.050
	2.70	5	.406	.400	.300	.244	.372	.250	.378	.262	.296	.342	.350
		7	.452	.504	.318	.294	.390	.288	.400	.300	.320	.342	.350
		10	.544	.606	.360	.386	.452	.386	.444	.374	.386	.418	.416
		15	.614	.704	.408	.422	.476	.416	.482	.386	.398	.432	.452
		22	.706	.822	.454	.482	.506	.490	.520	.438	.454	.476	.472
		30	.756	.874	.538	.554	.564	.568	.578	.508	.536	.548	.546
Strongly leptokurtotic/ strongly leptokurtotic	1.00	5	.096	.098	.054	.038	.078	.038	.076	.064	.068	.082	.072
		7	.122	.150	.074	.060	.090	.054	.090	.082	.074	.084	.086
		10	.136	.190	.068	.056	.078	.054	.068	.062	.062	.072	.084
		15	.136	.192	.060	.060	.072	.062	.076	.062	.054	.068	.066
		22	.132	.220	.046	.038	.054	.048	.054	.040	.036	.048	.046
		30	.146	.258	.052	.042	.042	.036	.040	.032	.040	.040	.036

(*cotninued*)

APPENDIX. (continued)

Distribution	F	n_y	CV^2	CL1	CL2	mlog	mlog2	mratio	mratio2	TG	R^2	Kl	FK
Strongly leptokurtotic/ strongly leptokurtotic (continued)	2.70	5	.366	.402	.240	.296	.328	.194	.336	.240	.234	.286	.298
		7	.448	.514	.314	.274	.368	.274	.378	.288	.258	.314	.346
		10	.532	.610	.386	.334	.394	.336	.412	.328	.280	.340	.362
		15	.646	.724	.446	.370	.432	.390	.444	.374	.318	.396	.398
		22	.708	.816	.478	.448	.454	.452	.488	.396	.356	.432	.446
		30	.774	.852	.526	.488	.496	.502	.514	.436	.450	.466	.480
Moderately skewed/ moderately skewed	1.00	5	.024	.050	.004	.002	.050	.002	.046	.030	.072	.052	.052
		7	.018	.048	.004	.002	.026	.002	.024	.030	.086	.042	.016
		10	.004	.032	.002	.000	.004	.000	.008	.016	.066	.018	.012
		15	.000	.008	.000	.000	.000	.000	.000	.004	.070	.004	.000
		22	.000	.000	.000	.000	.000	.000	.000	.000	.068	.000	.000
		30	.000	.000	.000	.000	.000	.000	.000	.000	.064	.000	.000
	2.70	5	.358	.402	.208	.212	.370	.216	.374	.232	.296	.370	.354
		7	.452	.530	.250	.298	.444	.294	.446	.294	.320	.416	.430
		10	.550	.642	.286	.406	.506	.406	.500	.360	.386	.476	.482
		15	.646	.754	.312	.508	.594	.508	.596	.480	.398	.554	.566
		22	.736	.840	.392	.618	.650	.620	.660	.554	.454	.644	.634
		30	.796	.896	.416	.690	.700	.688	.700	.630	.536	.658	.694
Strongly skewed/ strongly skewed	1.00	5	.076	.094	.026	.030	.076	.022	.082	.054	.068	.106	.078
		7	.086	.120	.026	.036	.072	.026	.068	.054	.074	.082	.066
		10	.100	.138	.024	.038	.058	.040	.056	.060	.062	.066	.052
		15	.084	.146	.026	.026	.038	.026	.042	.040	.054	.042	.040
		22	.090	.178	.024	.034	.040	.042	.042	.050	.036	.050	.038
		30	.100	.208	.022	.040	.044	.036	.040	.056	.040	.054	.038
	2.70	5	.350	.386	.186	.196	.350	.202	.358	.204	.276	.338	.332
		7	.434	.496	.270	.290	.422	.294	.400	.308	.298	.428	.394
		10	.540	.610	.308	.376	.454	.382	.468	.354	.318	.482	.436
		15	.614	.696	.352	.478	.524	.470	.530	.454	.366	.516	.498
		22	.708	.804	.386	.570	.596	.572	.596	.562	.394	.576	.576
		30	.786	.876	.454	.640	.650	.658	.678	.606	.450	.644	.642
Strongly skewed/ moderately bimodal	1.00	5	.042	.074	.016	.018	.068	.022	.078	.058	.074	.150	.090
		7	.054	.086	.012	.040	.088	.042	.092	.118	.136	.162	.108
		10	.048	.104	.010	.070	.104	.070	.106	.178	.152	.152	.094
		15	.050	.110	.004	.096	.136	.098	.144	.220	.194	.130	.108
		22	.068	.152	.006	.144	.156	.148	.160	.264	.232	.120	.120
		30	.074	.200	.016	.198	.214	.206	.216	.368	.298	.122	.118
	2.70	5	.506	.580	.242	.288	.494	.278	.532	.338	.420	.510	.494
		7	.602	.686	.298	.446	.612	.424	.638	.452	.552	.628	.614
		10	.682	.762	.330	.616	.694	.622	.718	.610	.660	.716	.682
		15	.764	.842	.404	.756	.814	.754	.820	.762	.800	.794	.784
		22	.824	.896	.468	.882	.894	.888	.900	.882	.912	.856	.852
		30	.872	.944	.502	.938	.940	.936	.938	.926	.956	.870	.882
Normal/strongly bimodal	1.00	5	.012	.026	.000	.000	.032	.002	.030	.014	.016	.044	.104
		7	.006	.028	.000	.002	.040	.006	.060	.032	.036	.044	.098
		10	.012	.080	.000	.070	.080	.080	.086	.126	.082	.036	.054
		15	.024	.074	.000	.052	.104	.062	.110	.090	.092	.024	.128
		22	.032	.104	.000	.154	.162	.148	.162	.178	.128	.036	.106
		30	.034	.122	.002	.174	.184	.194	.198	.204	.166	.024	.126
	2.70	5	.486	.524	.092	.296	.534	.294	.532	.306	.356	.424	.532
		7	.580	.656	.090	.460	.608	.458	.618	.418	.488	.512	.598
		10	.664	.764	.098	.604	.688	.608	.712	.534	.594	.578	.684
		15	.782	.876	.120	.766	.830	.772	.826	.688	.756	.672	.828
		22	.844	.942	.134	.874	.894	.876	.892	.806	.842	.756	.882
		30	.876	.958	.150	.928	.938	.928	.938	.884	.900	.800	.906
Strongly leptokurtotic/ strongly skewed	1.00	5	.092	.110	.036	.048	.100	.042	.098	.096	.090	.088	.080
		7	.096	.132	.036	.050	.098	.048	.106	.132	.120	.078	.078
		10	.098	.156	.048	.064	.094	.060	.098	.158	.130	.064	.064
		15	.104	.162	.046	.058	.086	.058	.088	.172	.132	.066	.060
		22	.126	.202	.052	.104	.112	.102	.112	.232	.180	.070	.068
		30	.148	.250	.052	.108	.114	.120	.126	.254	.182	.070	.066
	2.70	5	.356	.390	.254	.228	.370	.224	.366	.306	.324	.334	.306
		7	.452	.510	.312	.304	.440	.302	.446	.412	.410	.388	.388
		10	.510	.600	.344	.434	.506	.438	.502	.530	.492	.406	.432
		15	.616	.724	.400	.530	.608	.544	.606	.656	.618	.496	.522
		22	.700	.802	.450	.640	.672	.662	.692	.746	.710	.548	.582
		30	.758	.874	.522	.722	.734	.746	.760	.798	.778	.570	.644

(*continued*)

APPENDIX. (continued)

Distribution	F	n_y	CV^2	CL1	CL2	mlog	mlog2	mratio	mratio2	TG	R^2	Kl	FK
Moderately bimodal/ normal	1.00	5	.040	.062	.006	.006	.068	.010	.068	.030	.046	.082	.062
		7	.038	.058	.000	.008	.046	.014	.042	.016	.032	.072	.080
		10	.032	.032	.000	.010	.032	.014	.030	.012	.024	.064	.058
		15	.024	.066	.000	.010	.020	.012	.022	.014	.014	.044	.044
		22	.016	.088	.000	.010	.012	.012	.014	.006	.008	.040	.030
		30	.020	.088	.000	.012	.012	.012	.014	.008	.012	.036	.036
	2.70	5	.440	.484	.174	.260	.438	.270	.432	.244	.304	.430	.454
		7	.500	.570	.198	.338	.468	.334	.470	.260	.336	.498	.550
		10	.578	.686	.220	.390	.510	.404	.496	.282	.372	.584	.560
		15	.692	.814	.250	.476	.564	.460	.578	.342	.450	.660	.654
		22	.798	.914	.262	.586	.642	.600	.632	.422	.538	.724	.702
		30	.828	.944	.266	.628	.664	.642	.668	.452	.576	.758	.742
Strongly bimodal/ moderately skewed	1.00	5	.056	.076	.002	.004	.030	.004	.026	.010	.020	.100	.062
		7	.056	.076	.008	.008	.028	.006	.024	.012	.016	.116	.062
		10	.044	.078	.000	.006	.010	.004	.012	.010	.004	.112	.028
		15	.038	.086	.000	.000	.006	.002	.006	.012	.004	.080	.020
		22	.032	.076	.000	.002	.002	.000	.002	.016	.000	.072	.006
		30	.026	.088	.000	.000	.000	.002	.004	.014	.002	.066	.002
	2.70	5	.372	.424	.108	.152	.312	.148	.314	.140	.182	.374	.330
		7	.444	.518	.118	.200	.360	.182	.332	.162	.202	.454	.394
		10	.526	.624	.112	.280	.360	.268	.342	.174	.216	.544	.396
		15	.626	.778	.114	.316	.398	.290	.390	.246	.290	.590	.452
		22	.738	.866	.106	.392	.440	.378	.416	.268	.304	.626	.492
		30	.854	.946	.116	.444	.482	.428	.454	.300	.350	.660	.536
Moderately skewed/ moderately bimodal	1.00	5	.026	.042	.004	.010	.046	.008	.058	.048	.048	.092	.084
		7	.030	.050	.000	.016	.066	.028	.074	.072	.064	.094	.088
		10	.040	.062	.000	.048	.074	.056	.080	.134	.106	.084	.070
		15	.036	.078	.000	.056	.088	.066	.102	.142	.098	.074	.074
		22	.038	.112	.000	.104	.120	.100	.126	.200	.156	.080	.082
		30	.044	.142	.000	.138	.142	.138	.142	.232	.162	.084	.072
	2.70	5	.472	.540	.200	.268	.514	.272	.510	.316	.388	.466	.490
		7	.572	.630	.258	.436	.606	.432	.620	.408	.500	.560	.592
		10	.650	.750	.298	.586	.674	.592	.684	.596	.650	.646	.672
		15	.738	.844	.350	.728	.792	.734	.800	.734	.782	.754	.778
		22	.806	.924	.388	.870	.878	.854	.884	.860	.882	.808	.852
		30	.850	.942	.444	.926	.932	.926	.936	.914	.938	.848	.886

[1] CL1, Cope and Lacy (1992) test; CL2, Cope and Lacy (1994) test; CV^2, CV ratio test; F, CV_y^2/CV_x^2; FK, Fligner-Killeen test; Kl, Klotz test; mlog, median log test; mlog2, median log test, median values discarded; mratio, median ratio test; mratio2, median ratio test, median values discarded; n_y, fossil n; R^2, squared ranks test; TG, Talwar-Gentle test.

LITERATURE CITED

Bailer AJ. 1989. Testing variance equality with randomization tests. J Stat Comp Simul 31:1–8.

Box GEP. 1953. Non-normality and tests on variances. Biometrika 40:318–335.

Brown MB, Forsythe AB. 1974. Robust tests for the equality of variances. J Am Stat Assoc 69:364–367.

Calcagno JM, Cope DA, Lacy MG, Tobias PV. 1997. Is *A. africanus* the only hominid species in Sterkfontein Member 4? Am J Phys Anthropol 24(Suppl):86 (abstract).

Chambers JM, Cleveland WS, Kleiner B, Tukey PA. 1983. Graphical methods for data analysis. Belmont, CA: Wadsworth.

Coffing KE, Teaford MF. 1993. Range-based analyses of molar size variation in early *Homo*. Am J Phys Anthropol 18(Suppl.):68–69 (abstract).

Cole TM, Smith FH. 1987. An odontometric assessment of variability in *Australopithecus afarensis*. Hum Evol 3:221–234.

Conover WJ. 1980. Practical nonparametric statistics, 2nd ed. New York: John Wiley & Sons.

Conover WJ, Iman RL. 1978. Some exact tables for the squared ranks test. Comm Stat B—Simulation Comp 7:491–513.

Conover WJ, Johnson ME, Johnson ME. 1981. A comparative study of tests for homogeneity of variances, with application to the outer continental shelf bidding data. Technometrics 23:351–361.

Cope DE. 1993. Measures of variation as indicators of multiple taxa in samples of sympatric *Cercopithecus* species. In: Kimbel WH, Martin LB, editors. Species, species concepts and primate evolution. New York: Plenum Press. p 211–237.

Cope DE, Lacy MG. 1992. Falsification of a single species hypothesis using the coefficient of variation. A simulation approach. Am J Phys Anthropol 89:359–378.

Cope DE, Lacy MG. 1994. Testing single species hypotheses using the combined referent CV: applications to fossil hominoid dental samples. Am J Phys Anthropol 18(Suppl.):70 (abstract).

Cope DE, Lacy MG. 1995. Comparative application of the coefficient of variation and range-based statistics for assessing the taxonomic composition of fossil samples. J Hum Evol 29:549–576.

Donnelly SM. In press. Range-based measures of variation are not immune to Type 1 errors. Am J Phys Anthropol.

Fligner MA, Killeen TJ. 1976. Distribution-free two sample tests for scale. J Am Stat Assoc 71:210–213.

Foote M. 1993. Human cranial variability: a methodological comment. Am J Phys Anthropol 90:377–379.
Fuller K. 1996. Analysis of the probability of multiple taxa in a combined sample of Swartkrans and Kromdraai dental material. Am J Phys Anthropol 101:429–439.
Gelvin BR, Albrecht GH, Miller JMA. 1997. The hierarchy of craniometric variation among gorillas. Am J Phys Anthropol 24(Suppl.):117 (abstract).
Godfrey LM, Lyon SK, Sutherland MR. 1993. Sexual dimorphism in large-bodied primates: the case of the subfossil lemurs. Am J Phys Anthropol 90:315–334.
Good P. 1993. Permutation tests. New York: Springer-Verlag.
Grine FE. 1988. New craniodental fossils of *Paranthropus* from the Swartkrans Formation and their significance in "robust" australopithecine evolution. In: Grine FE, editor. Evolutionary history of the "robust" Australopithecines. New York: Aldine de Gruyter. p 223–243.
Grine FE, Daegling DJ. 1993. New mandible of Paranthropus robustus from Member 1, Swartkrans Formation, South Africa. J Hum Evol 24:319–333.
Grine FE, Strait DS. 1994. New hominid fossils from Member I "Hanging Remnant," Swartkrans Formation, South Africa. J Hum Evol 26:57–75.
Groves CP. 1989. A theory of primate and human evolution. Oxford: Clarendon Press.
Hochberg Y, Tamhane AC. 1987. Multiple comparison procedures. New York: John Wiley and Sons.
Johanson DC, White TD, Coppens Y. 1982. Dental remains from the Hadar Formation, Ethiopia: 1974–1977 collections. Am J Phys Anthropol 57:545–603.
Josephson SC, Juell KE, Rogers AR. 1996. Estimating sexual dimorphism by method-of-moments. Am J Phys Anthropol 100:191–206.
Kanji GK. 1993. 100 statistical tests. Newbury Park, CA: Sage Publications.
Kay RF. 1982. Sexual dimorphism in Ramapithecinae. Proc Natl Acad Sci U S A 79:209–212.
Kelley J. 1986. Species recognition and sexual dimorphism in *Proconsul* and *Rangwapithecus*. J Hum Evol 15:461–495.
Kelley J. 1993. Taxonomic implications of sexual dimorphism in *Lufengpithecus*. In: Kimbel WH, Martin LB, editors. Species, species concepts and primate evolution. New York: Plenum Press. p 429–458.
Kelley J, Xu Q. 1991. Extreme sexual dimorphism in a Miocene hominoid. Nature 352:151–153.
Kimbel WH, White TD. 1988. Variation, sexual dimorphism and the taxonomy of *Australopithecus*. In: Grine FE, editor. Evolutionary history of the "robust" Australopithecines. New York: Aldine de Gruyter. p 175–192.
Kimbel WH, White TD, Johanson DC. 1985. Craniodental morphology of the hominids from Hadar and Laetoli: Evidence of "*Paranthropus*" and *Homo* in the mid-Pliocene of eastern Africa? In: Delson E, editor. Ancestors: the hard evidence. p 120–137.
Klockars AJ, Sax G. 1986. Multiple comparisons. Sage University paper series on quantitative applications in the social sciences, 07–061. Beverly Hills, CA: Sage Publications.
Klotz J. 1962. Nonparametric tests for scale. Ann Math Stat 32:498–512.
Koopman LH. 1981. An introduction to contemporary statistics. Boston: Duxbury.
Lee JJ, Tu ZN. 1997. A versatile one-dimensional distribution plot: the BLiP plot. American Statistician 51:353–358.
Levene H. 1960. Robust tests for equality of variances. In: Olkin I, editor. Contributions to probability and statistics. Stanford: Stanford University Press. p 278–292.
Lewontin RC. 1966. On the measurement of relative variability. Syst Zool 15:141–142.
Mahler PE. 1973. Metric variation in the pongid dentition. PhD dissertation. Ann Arbor: University of Michigan.
Marascuilo LA, McSweeney M. 1977. Nonparametric and distribution-free methods for the social sciences. Belmont, CA: Wadsworth.
Martin CG, Games PA. 1977. ANOVA tests for homogeneity of variance: nonnormality and unequal samples. J Educ Stat 3:187–206.
Martin L. 1991. Teeth, sex and species. Nature 352:111–112.
Martin L, Andrews P. 1984. The phyletic position of *Graecopithecus freybergi* Koenigswald. Cour Forsch Inst Senk 69:25–40.
Martin L, Andrews P. 1993. Species recognition in middle Miocene hominoids. In: Kimbel WH, Martin LB, editors. Species, species concepts and primate evolution. New York: Plenum Press. p 393–427.
Miller JMA. 1994. *Homo habilis:* is the degree of craniofacial variation excessive? Am J Phys Anthropol 18(Suppl):147 (abstract).
Miller JMA. 1995. Craniofacial variation in *Homo habilis*: a multivariate comparison of KNM-ER 1470 and KNM-ER 1813. Am J Phys Anthropol 20(Suppl):154 (abstract).
Miller JMA, Albrecht GH. 1997. An hierarchical analysis of craniofacial variation in *Homo habilis* using a *Gorilla* analog. Am J Phys Anthropol 24(Suppl):170 (abstract).
Miller JMA, Albrecht GH, Gelvin BR. 1998. A hierarchical analysis of craniofacial variation in *Homo habilis* compared to a modern human analog. Am J Phys Anthropol 26(Suppl):162 (abstract).
Miller RG. 1968. Jackknifing variances. Ann Math Stat 39:567–582.
Mood AM. 1954. On the asymptotic efficiency of certain non-parametric two-sample tests. Ann Math Stat 25:514–533.
Myers JL, Well AD. 1995. Research design and statistical analysis. Hillsdale, NJ: Lawrence Erlbaum Associates.
O'Brien RG. 1978. Robust techniques for testing heterogeneity of variance effects in factorial designs. Psychometrika 43:327–342.
Pan Y, Waddle DM, Fleagle JG. 1989. Sexual dimorphism in *Laccopithecus robustus*, a late Miocene hominoid from China. Am J Phys Anthropol 79:137–158.
Pearson ES. 1926. A further note on the distribution of range in samples taken from a normal population. Biometrika 18:173–194.
Plavcan JM. 1993. Catarrhine dental variability and species recognition in the fossil record. In: Kimbel WH, Martin LB, editors. Species, species concepts and primate evolution. New York: Plenum Press. p 239–263.
Plavcan JM. 1994. Comparison of four simple methods for estimating sexual dimorphism in fossils. Am J Phys Anthropol 94:465–476.
Schultz B. 1983. On Levene's test and other statistics of variation. Evol Theory 6:197–203.
Schultz B. 1985. Levene's test for relative variation. Syst Zool 34:449–456.
Sokal RR, Rohlf FJ. 1981, Biometry, 2nd ed. New York: WH Freeman.
Stringer CB. 1986. The credibility of *Homo habilis*. In: Wood BA, Martin LB, Andrews P, editors. Major topics in primate and human evolution. Cambridge: Cambridge University Press. p 266–294.
Talwar PP, Gentle JE. 1977. A robust test for the homogeneity of variances. Comm Stat A—Theor Meth 6:363–369.

Teaford MF, Walker A, Mugasi GS. 1993. Species discrimination in *Proconsul* from Rusinga Islands, Kenya. In: Kimbel WH, Martin LB, editors. Species, species concepts and primate evolution. New York: Plenum Press. p 373–393.

Toothaker LE. 1993. Multiple comparison procedures. Sage University paper series on quantitative applications in the social sciences, 07–089. Beverly Hills, CA: Sage Publications.

Van Valen L. 1977. The statistics of variation. Evol Theory 4:33–43.

Vinyard P. 1997. Postcranial variation in extant hominoids with implications for interpreting hominid fossil assemblages. J Hum Evol 32:A24 (abstract).

Walker A, Teaford MF, Martin L, Andrews P. 1993. A new species of *Proconsul* from the early Miocene of Rusinga/Mfango Islands, Kenya. J Hum Evol 25:43–56.

Walpole RE, Myers RH. 1989. Probability and statistics for engineers and scientists. New York: Macmillan Publishing.

White TD. 1977. New fossils hominids from Laetolil, Tanzania: 1976–1979 specimens. Am J Phys Anthropol 46:197–230.

White TD. 1980. Additional fossil hominids from Laetoli, Tanzania: 1976–1979 specimens. Am J Phys Anthropol 53:487–504.

Wolpoff MH. 1978. Analogies and interpretation in palaeoanthropology. In: Jolly CJ, editor. Early hominids of Africa. London: Duckworth. p 461–503.

Wolpoff MH. 1992. Levantines and Londoners. Science 255:142.

Wood BA. 1991. Koobi Fora Research Project, vol 4. Hominid cranial remains. Oxford: Clarendon Press.